PCT/JP2004/010250

12.7.2004

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００３年１０月　７日

出願番号
Application Number: 特願２００３－３４８４３８

[ST. 10/C]: [ＪＰ２００３－３４８４３８]

出願人
Applicant(s): 株式会社リバース・プロテオミクス研究所

２００４年　８月１３日

特許庁長官
Commissioner,
Japan Patent Office

小川洋

出証番号　出証特２００４－３０７２４０７

【書類名】　特許願
【整理番号】　P03-0895
【提出日】　平成15年10月 7日
【あて先】　特許庁長官　殿
【国際特許分類】　G06F159/00
【発明者】
　【住所又は居所】　千葉県木更津市かずさ鎌足二丁目６番地７　株式会社リバース・プロテオミクス研究所内
　【氏名】　飛田　基
【発明者】
　【住所又は居所】　千葉県木更津市かずさ鎌足二丁目６番地７　株式会社リバース・プロテオミクス研究所内
　【氏名】　西川　哲夫
【発明者】
　【住所又は居所】　千葉県木更津市かずさ鎌足二丁目６番地７　株式会社リバース・プロテオミクス研究所内
　【氏名】　堀内　健
【発明者】
　【住所又は居所】　千葉県木更津市かずさ鎌足二丁目６番地７　株式会社リバース・プロテオミクス研究所内
　【氏名】　根本　昌
【発明者】
　【住所又は居所】　千葉県木更津市かずさ鎌足二丁目６番地７　株式会社リバース・プロテオミクス研究所内
　【氏名】　荒木　健司
【特許出願人】
　【識別番号】　501260082
　【氏名又は名称】　株式会社　リバース・プロテオミクス研究所
【代理人】
　【識別番号】　100091096
　【弁理士】
　【氏名又は名称】　平木　祐輔
【選任した代理人】
　【識別番号】　100105463
　【弁理士】
　【氏名又は名称】　関谷　三男
【選任した代理人】
　【識別番号】　100102576
　【弁理士】
　【氏名又は名称】　渡辺　敏章
【選任した代理人】
　【識別番号】　100103931
　【弁理士】
　【氏名又は名称】　関口　鶴彦
【手数料の表示】
　【予納台帳番号】　015244
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　特許請求の範囲　1
　【物件名】　明細書　1

【物件名】　図面　1
【物件名】　要約書　1

【書類名】特許請求の範囲
【請求項１】
二つの生体関連事象間の相関データあるいは該相関データとそれぞれの事象の特徴データを行列形式で表示する可視化方法において、同一種類または異なる種類の生体関連事象間の相関データあるいは該相関データと各生体関連事象の特徴データを、所望する表示データのデータ数に応じて、予め用意された（ａ）複数のデータ表示形式から手動または自動的に選択された一つの形式と、（ｂ）複数のデータ要約度から手動または自動的に選択された一つの要約度に基づいて画面表示することを特徴とする生体関連事象間相関データの可視化方法。
【請求項２】
前記（ａ）複数のデータ表示形式として、（Ａ）一対の事象間の相関データを一つの表示データ単位とする表形式のデータ表示形式、（Ｂ）事象をクラスタリングした結果得られたクラスター間の相関データを一つの表示データ単位とする表形式のデータ表示形式、及び（Ｃ）相関データの集合を統計処理した結果を一つの表示データ単位とするデータの表示形式から選択される表示形式を用いることを特徴とする請求項１に記載の可視化方法。
【請求項３】
前記（ｂ）複数のデータ要約度として、データフィールドの表示または非表示、文字型のデータフィールド中のデータの短縮、及び数値型データフィールド中のデータの短縮から選択される要約方法を用いることを特徴とする請求項１に記載の可視化方法。
【請求項４】
前記文字型のデータフィールド中のデータの短縮が、階層構造を有する文字情報から該階層の一部分を抽出する操作、文字データ中からあらかじめ登録されているキーワードを抽出する操作、及び文字データを一つの記号や文字または色彩で対応させる操作からなることを特徴とする請求項３に記載の可視化方法。
【請求項５】
前記数値型のデータフィールド中のデータの短縮が、数値を任意の有効数字で丸める操作、数値の指数部分のみを取り出す操作、及び一定範囲の数値を色彩で対応させる操作からなることを特徴とする請求項３に記載の可視化方法。
【請求項６】
画面表示形式とデータの要約度の自動的な選択方法として、画面表示すべき相関データのエントリー数及びあらかじめ指定された情報表示領域と情報表示単位のサイズに応じて、最大の情報量を与えるデータ表示形式とデータ要約度の組を選択することを特徴とする請求項１に記載の可視化方法。
【請求項７】
前記生体関連事象間の相関データが、低分子化合物とタンパク質の相互作用であることを特徴とする請求項１に記載の可視化方法。
【請求項８】
請求項１～請求項７に記載の可視化方法をコンピューターに実行させるためのプログラムを記録したコンピューター読み取り可能な記録媒体。
【請求項９】
前記プログラムが、同一種類または異なる種類の生体関連事象間の相関データ、あるいは該相関データとそれぞれの事象の特徴データをメモリ領域に形成する処理、所望する表示データのデータ数に応じて、予め用意された（ａ）複数のデータ表示形式から手動または自動的に選択された一つの形式と、（ｂ）複数のデータ要約度から手動または自動的に選択された一つの要約度により画面表示する処理を含むことを特徴とする請求項８に記載の記録媒体。

【書類名】明細書
【発明の名称】生体関連事象間の相関データの可視化方法
【技術分野】
【０００１】
本発明は、生体関連事象間の相関データ、特に蛋白質、低分子化合物、ＤＮＡ等の生体内物質間の相互作用情報や遺伝子の発現プロファイル等の視覚化方法に関する。また、本発明は、上記方法を取り入れたグラフィカルユーザーインターフェース、及び可視化システムに関する。
【背景技術】
【０００２】
ヒトゲノム計画の完了に伴い、遺伝子配列、さらにはそこにコードされたタンパク質配列情報が網羅的に集積されてきている。現在、新しい診断方法や新薬の創出を目指して、これらの配列情報やタンパク質を用いた機能解析が、盛んに行われている。タンパク質の機能を調べる上で、タンパク質－タンパク質相互作用を知ることには、非常に重要な意味がある。それは、他の生体内物質との相互作用が、タンパク質の機能そのものであるからである。タンパク質－タンパク質相互作用以外にも、遺伝子のライブラリー毎の発現プロファイルやタンパク質－低分子化合物相互作用などのように、二つの物質、一般化していえば二つの事象間の相関関係情報は、生体内物質のシステムとしての機能の解明に寄与すると考えられ、近年大規模なデータ収集が行われ始めている。データ量が増大すればするほど、データ全体を概観しそこから特徴を抽出することが困難になってくると考えられる。また、データ量が増大すれば、個別データの詳細な参照が多数必要となって、個別サイトの観察が頻繁になるという問題があった。そこで、大量の相関データから、その中に埋もれている情報を有効に抽出するために、情報可視化方法の重要性が増大している。
【０００３】
大量の相関データの可視化方法として、一方の事象を行、もう一方の事象を列にとった行列を考え、この行列の交差するセル内に二つの事象間の相関データを記述する表示方法がある。発現プロファイルでは、行列のセル内に発現強度に応じた色彩を表示する方法が、一般に用いられている。タンパク質－タンパク質相互作用の可視化においても、行列のセル内に相互作用に応じた色彩あるいは濃淡を表示する方法が行われている。タンパク質－低分子化合物相互作用の可視化においても、行列のセル内に相互作用に応じた「＋＋」、「＋」などの定性的情報を表示する方法が行われている（下記特許文献１）。
【０００４】
二つの事象間の相関関係情報を行列で表示する方法においては、行列上の相関データのパターンをもとにクラスタリングを行うことが一般的に行われている。得られたクラスター内の事象がどういう事象かを解析することによって、相関情報と各事象の特徴間の関連がわかる。また同様に、各事象の特徴によって事象のソートを行い、得られた相関情報パターンと事象の特徴を比較することによって、相関情報と各事象の特徴間の関連がわかる。このように、行列による相関データの可視化方法においては、相関情報のパターンと各事象の特徴を両方観察できることが重要である。
【０００５】
従って、情報閲覧の有効な方法として、まず、データ数の規模が大きい相関データに対する行列表示を行い、相関データパターンによるクラスタリングや各事象の特徴による事象のソート等によって特徴的なパターンを同定する。その後、同定したパターンの構成要素に関する特徴量や相互作用情報の詳細情報にアクセスしていくことによって、得られたパターンの持つ意味について考察していくことが可能になる。さらに、上述したクラスタリングやソートと異なる方式でクラスタリングやソートを再度行い、得られた相関データパターンの全体を観察し、その中で先に注目した個別の相互作用と事象がどのようなクラスターに属しているかを調べることによって、新たな発見に繋がる可能性が生じる。このように、大量の相関データ行列表示と個別の相関データ表示との間で交互に行き来を繰り返すことによって、相関データに関する新しい知識の発見が可能になると考えられる。

【０００６】
しかしながら、従来の行列による相関データの可視化方法においては、データ数の規模が大きく変動した際に、規模に応じた適切な情報が得られないという問題があった。例えば、画面の画素数が縦横1,000ピクセル×1,000ピクセル程度（大きさで言えば30cm×30cm）程度であるとしよう。データ規模が数十個～百個のオーダーの場合は、一つのセルあたりの画素数は10～数十ピクセル×10～数十ピクセルで、大きさにして数mm^2～$1cm^2$程度になり、色彩あるいは濃淡のパターンと各データポイント一つ一つが同時に観察可能である。
【０００７】
しかし、データ規模が数百個以上に増大した場合は、一つのセルあたりの画素数は数ピクセル×数ピクセル以下で、一つのセルの大きさは$1mm^2$以下程度になる。この場合は、セルが小さすぎてパターン情報が複雑になると同時に、セル一個一個の認識が困難になってくる。また、描画時間がかかるという問題も生じてくる。このように、データ規模が数百個以上に増大した場合には、一定数のセルあるいはクラスターに対応した複数のセルをまとめて一つの相関データを記載するパターンの粗視化を選択することで、一つのセルのサイズが数mm～1cm×数mm～1cm程度になり、相関データパターンと各データポイント一つ一つが同時に観察可能になる。従来は、この操作をユーザーがマニュアルによって実施する必要があり、手間がかかっていた。
【０００８】
逆に、行ないし列の規模が数十個以下に減少した場合は、一つのセルあたりの画素数が数十ピクセル×数十ピクセル以上で、一つのセルの大きさにして数cm^2以上と大きいにも関わらず、セル当りの情報量が色彩で表現される程度の情報量のままであるため、画面全体から得られる情報量が減少してくる。画面全体から得られる情報量を増加させるために、個々のセルに関する情報を参照しようとすれば、個々のセル毎に別の情報ソースにアクセスする必要が生じてくる。この場合、相関データパターンと、パターンを構成する複数のセルに関する情報を同時に参照することが困難であり、また手間も大きかった。
【０００９】
【特許文献１】国際公開公報:WO 02/23199 A2
【非特許文献１】Advanced Drug Delivery Reviews, 23 (1997) 3-25
【発明の開示】
【発明が解決しようとする課題】
【００１０】
以上述べてきたように、二つの事象間の相関データを行列形式で表示する可視化方法において、相関データパターンとパターンを構成する複数のセルに関する情報を同時に観察するためには、相関データ規模の大小によって、相関データパターンの粗視化（クラスタリング等によって複数のセルをまとめて要約する作業）や、セル毎の情報の他ソースへのアクセス等の作業を実施する必要があった。しかも、従来の方法では、これらの作業はマニュアルによって行わなければならなかった。従来の技術で述べたように、大量の相関データから有効な知識を発見するためには、相関データの全体としての観察と少数データの詳細な観察を交互に繰り返す作業が必要である。従来のマニュアルによる方法は、この繰り返し作業を行う際の効率が非常に低かった。
【００１１】
本発明は、二つの事象間の相関データを行列形式で表示する可視化方法において、相関データパターンとパターンを構成する複数のセルに関する情報を、データ数の規模の変動に応じて適切な形式で、同時に観察する手段を提供することを目的とする。
【課題を解決するための手段】
【００１２】
前記目的を達成するため、本発明による二つの事象間の相関データを行列形式で表示する画面表示システムは、データ数の規模の変動に応じて、予め用意された複数の単位相関データあたりのデータの集積度が異なるデータ表示形式の中から一つを自動的に選択し、

また、個々のセルに関する情報（相関や各事象に関する情報）について予め用意された複数の要約度が異なる表示方法の中から一つを自動的に選択して、相関データと個々のセルに関する情報を表示することを特徴とする。

【００１３】

二つの事象間の相関データの典型例としては、一方の事象は蛋白質、もう一方の事象は低分子化合物、事象間の相関データは蛋白質－低分子化合物間の相互作用の強さである。また、両方の事象共に蛋白質で、事象間の相関データは蛋白質－蛋白質間の相互作用の強さ、あるいは蛋白質間の配列類似性であってもよい。さらに、一方の事象は遺伝子、もう一方の事象は遺伝子が由来するｃＤＮＡライブラリーであって事象間の相関データは遺伝子のｃＤＮＡライブラリー毎の発現強度であってもよい。また、両方の事象共に低分子化合物で、事象間の相関データは低分子化合物間の構造類似性や薬効上又は副作用上の相互作用であってもよい。

【００１４】

本発明による画面表示方法は、データ表示形式として、（Ａ）相関データの要素そのもの、例えば低分子化合物とタンパク質の結合定数、を画面表示データ単位とする表示形式（個別データ表示形式と呼ぶ）、（Ｂ）複数の相互作用データのまとまりを画面表示データ単位とする表示形式（相関データのパターンや事象の特徴に基づくクラスタリングから得られたクラスターを、複数の相互作用データのまとまりとする。そこで、クラスター表示形式と呼ぶ）、（Ｃ）複数の相関データの統計値を画面表示データ単位とする表示形式（統計表示形式と呼ぶ）の三つを有することを特徴とする。相関データの統計値とは、クラスターの数そのものや、クラスターの各要素について別のデータソースから得られる関連情報の数などをいう。

【００１５】

本発明による画面表示方法は、個々のセルに関する情報（相関や各事象に関する情報）の表示方法として、情報量に依存して設定された複数の要約度に従った表示方法を有することを特徴とする。要約度は、一つの事象を表現する際の情報量が小さいほど高い値として定義される。

【００１６】

本発明によって定義される複数の要約度は、以下のとおりである。データフィールドに格納されている意味的に重複しない全ての情報を画面に出力するとき、データは要約されていないので、データの要約度は０であるとする。異なる種類のデータフィールドに対して、それぞれ複数の要約度に対応するデータのフォーマットを定義しておく。例えば、指数部分を含む実数データの表示において、
要約度０ではフィールド値そのものを表示、
要約度１では指数部分のみを表示、
要約度２では指数部分の値を五つのクラスターに分類し、クラスターに対応する色で情報を表示、
要約度３ではある閾値以上のもののみ色をつけて表示、
とすることができる。また、階層構造を表している文字列データの表示において、
要約度０では階層構造のそれぞれの定義を階段状に表示、
要約度１では階層構造の最上層または最下層の定義のみを表示、
要約度２では階層構造の最上層または最下層に対応する情報をシンボルや色彩に射影して表示、
要約度３では階層構造の最上層の値に対応する色をつけて表示、
とすることができる。

【００１７】

本発明による画面表示方法は、データ数の規模の変動に応じて、上述した複数のデータ表示形式の中から一つを自動的に又は手動で選択するステップ、また上述した個々のセルに関する情報（相関や各事象に関する情報）の要約度の異なる複数の表示方法の中から一つを自動的に、あるいは手動で選択するステップ、及び選択したデータ表示形式と要約度

を用いて、相関データと各事象に関する情報を表示するステップ、とを含むことを特徴とする。

【００１８】

本発明によるデータ表示形式と要約度を自動的に選択する場合、画面に表示される情報量をユーザーが認識可能な最大の情報量付近の一定の値の近傍に留めるような選択を行うことを特徴とする。別の言い方をすれば、一つの画面に関連するすべての情報が表示されることを基準としてデータ表示形式と要約度を自動的に選択する。ただし、画面の少々のスクロールを許してよい。

【００１９】

以上のことを行うことによって、二つの事象間の相関データを行列形式で表示する可視化方法において、相関データ規模の大小によって、相関データパターンの粗視化や、セル毎の情報の他ソースへのアクセス等の作業をマニュアルで実施することなく、相関データパターンとパターンを構成する複数のセルに関する情報を、データ数の規模の変動に応じて自動的に選択された適切な形式で、同時に観察することが可能になる。これによって、相関データの全体としての観察と少数データの詳細な観察を交互に繰り返す作業を、従来のマニュアル操作に比べ大幅に効率的に実施することが可能になり、大量の相関データからの有効な知識の発見を効率的に行うことが可能になる。

【発明の効果】

【００２０】

二つの生体関連事象間の相関データを行列形式で表示する可視化方法において、本発明による可視化方法と、該可視化方法を実装したインターフェースを用いれば、相関データ規模の大小によって、相関データパターンの粗視化や、セル毎の情報の他ソースへのアクセス等の作業をマニュアルで実施することなく、相関データパターンとパターンを構成するセルに関する情報を、データ数の規模の変動に応じて自動的に選択された適切な表示形式と要約度で、同時に観察することが可能になる。これによって、表示すべきデータ数にかかわらず、個別セル内から得られる情報量を自動的に最大に保ちつつ、データの全体像の観察が可能になる。その結果、相関データの全体としての観察と少数データの詳細な観察を交互に繰り返す作業を、従来のマニュアルに比べ大幅に効率的に実施することが可能になり、大量の相関データからの有効な知識の発見を効率的に行うことが可能になる。

【発明を実施するための最良の形態】

【００２１】

以下、図面を参照して本発明の実施の形態を説明する。

【実施例１】

【００２２】

二つの事象間の相関関係として、蛋白質、低分子化合物、ＤＮＡ等の生体内物質間相互作用を考える。着目する二つの事象として「低分子化合物」と「タンパク質」間の相互作用データを扱う場合の実施例を、以下に説明する。ここで、相互作用データとは、Protein Data Bank (PDB, http://www.pdb.org)中に低分子化合物とタンパク質の複合体データがあるか、ないかという情報や、実験的に低分子化合物とタンパク質との間の結合の度合いを測定したデータである。タンパク質の特徴データとしては、各種外部データベースの情報や計算されたクラスタリング結果を持つ。例えば、SwissProt(http://www.expacy.ch/sprot)のIDや、アミノ酸配列相同性に基づいたクラスタリング結果、Gene Ontology(http://www.geneontology.org) に基づくアノテーション情報、溶媒への溶解度などである。低分子化合物の特徴データとしては、分子名、分子量、薬効分類、その他、電荷分布や親水・疎水性、立体構造、水素結合のドナー・アクセプター数、官能基の種類や数など様々な分子特性値を持つ。

【００２３】

まず、図１を用いてデータ可視化のフローチャートを説明する。ユーザー操作１０１はデータと実行するアクションを選択する部分である。アクションには、データ取得１０２とデータ処理１０３がある。データ取得には、各種検索条件による蛋白質－低分子化合物

相互作用データベース１０４からの検索によるデータ取得、表示画面上で指定された蛋白、あるいは低分子化合物に関連した各種相関関係テーブル１０５からのデータ取得がある。データ処理には、表示画面上で指定されたエントリーに対するクラスタリング等の処理や表示スケールの変更等の処理がある。取得、あるいは処理されたデータは表示データ１０６として扱われる。次に、表示データに対して、データの表示形式と要約度が決定される。データの表示形式と要約度は、表示データのデータ数に応じて、予め用意されたデータの表示形式と要約度決定ルール１０７に基づいて決定される。決定されたデータの表示形式と要約度に従い、データの画面表示１０８が行われる。各種相関関係テーブルとしては、タンパク質－タンパク質相互作用テーブル、タンパク質の発現プロファイルテーブル、低分子化合物－低分子化合物間の構造類似性や、薬効上または毒性上の相互作用テーブル等が考えられる。

【００２４】

本発明の要点である、「データの表示形式と要約度が、表示データのデータ数に応じて、予め用意されたデータの表示形式と要約度決定ルールに基づいて決定される」という点について、以下詳細に説明する。

【００２５】

まず、データの表示形式について説明する。図２に低分子化合物とタンパク質の相互作用データの画面表示例を示す。行列の縦方向に低分子化合物のラベル２０１、横方向にタンパク質のラベル２０２を並べ、行列部分２０３には実験的に測定されたタンパク質と低分子化合物の間の結合定数のうちある閾値より上のものに関して結合の強さ別に色の濃さを変えて表示している。また、化合物ラベルの左側には化合物の特徴量として分子量２０４を表示し、タンパク質ラベルの上側にはタンパク質の特徴量として、アルファヘリックスとベータストランドの数２０５と蛋白質相互の相同性に基づくクラスタリング情報２０６を表示している。

【００２６】

表形式で画面表示された相互作用データについては、相互作用データプロファイルに基づくクラスタリング、あるいは、タンパク質の特徴量や、低分子化合物の特徴量に基づくクラスタリングを行い、得られたクラスタリング情報に基づいてデータを並べ替えて表示することが可能である。

【００２７】

相互作用データを用いたクラスタリングは、例えば以下の方法によって行う。ひとつの低分子化合物C_iに着目して、それと各タンパク質P_jの相互作用強度プロファイルI_{ij}（j=1，．．．，N_p，N_pはタンパク質数）を考える。次に、全ての低分子化合物間で総当りの相互作用強度プロファイル間距離を計算する。低分子化合物C_iと低分子化合物C_k間の相互作用強度プロファイル間距離D_{ik}は、低分子化合物C_iとタンパク質P_j間の相互作用強度がI_{ij}とすれば、例えば以下の式によって計算される。

【００２８】

【数１】

$$D_{ik}=\sqrt{\Sigma\,(I_{ij}-I_{kj})^2}$$

上式中の和はj=1，．．．，N_pについてとる。

【００２９】

この式によって得られた総当りのD_{ik}に対して閾値を設けることによって、低分子化合物をクラスタリングすることが可能である。次に、ひとつのタンパク質P_iに着目して、それと各低分子化合物C_jの相互作用強度プロファイルI_{ij}（j=1，…，N_c，N_cは低分子化合

物数）を考える。低分子化合物の場合と同様にして、全てのタンパク質間で総当りの相互作用強度プロファイル間距離を計算することによって、タンパク質をクラスタリングすることが可能である。

上記のクラスタリングを実際に行った結果が、図３に示されている。

【００３０】

低分子化合物は３つ、タンパク質も３つのクラスターに分類され、その結果は低分子化合物のラベル上に低分子化合物クラスターA３０１、低分子化合物クラスターB３０２、低分子化合物クラスターC３０３として、またタンパク質のラベル上にタンパク質クラスターA３０４、タンパク質クラスターB３０５、タンパク質クラスターC３０６として色の濃さで識別表示されている。クラスター毎に相互作用データである結合定数の平均値が内部で計算され、クラスターは結合定数の平均によって上から下、左から右へ降順にソートされている。したがって、全体的な傾向として、マトリクス部分の左上のほうに結合定数の高い（色の濃い）セルが集まり、右下のほうには結合定数が低い又は閾値以下の結合しかないセルが集まっている。このような相互作用プロファイルに基づいたクラスタリングを行うことによって、特定の低分子化合物とタンパク質の組からなるクラスター３０７や、一つのタンパク質について特異的に相互作用をもつ多くの化合物を含むクラスター３０８などが視覚的に明らかになる。創薬研究への応用として、相互作用プロファイルに基づいて作成された低分子化合物のクラスターに共通する母核構造を抽出して、それを薬物の機能を担うファーマコフォアとして構造展開の種とするアプローチが可能である。

【００３１】

同様に、分子量をいくつかの区分に分けてクラスタリングしたり、タンパク質のアルファヘリックスとベータストランドの数をあるルールに従って分類したりすることが可能である。そして、分子量に基づくクラスター、アルファヘリックスとベータストランドの数に基づくクラスター、或いはあらかじめ計算されているアミノ酸配列の相同性に基づくクラスターのそれぞれについて表示データを並べ替えることができる。特に、ある特徴量についてデータを並べ替えた結果、特徴的な結合定数の色彩パターンが表れた場合には、その特徴量と結合定数が密接に関連していることを知ることができる。

【００３２】

図４に、データを低分子化合物側については分子量、タンパク質側についてはアミノ酸の相同性にもとづいてクラスタリングをし、クラスタリング結果によって表を並べ替えた結果を示す。低分子化合物は分子量によって分子量の比較的大きなクラスターA４０１、中程度の分子量を持つクラスターB４０２、分子量の比較的小さなクラスターC４０３に分類されており、データ全体は分子量について降順にソートされている。タンパク質は、アミノ酸配列の相同性に基づいてクラスター１、４０４とクラスター２，４０５が画面上に示されている。ここでは、クラスターBに属する低分子化合物が相互作用マトリクスの中では比較的相互作用が高い領域４０６と重なっているように見える。一方、アミノ酸の相同性に基づくクラスタリング結果と相互作用強度の間には明白に視認できるような相関は見当たらないようである。このように特徴量に関してクラスタリングを行い、その結果によってデータを並べ替えることによって、相互作用データをよく説明するような特徴量を発見できる可能性がある。低分子医薬品がもつ特徴量（分子特性）としてよく知られているものにChristopher A. Lipinski博士の“Rule of five”（上記非特許文献１）があるが、特徴量によるクラスタリング結果と相互作用データを同時に可視化することで、特定の実験データを説明する特徴量や、特定のタンパク質の標的となりうる低分子化合物が持つべき特徴量をルール化することも可能であると考えられる。

【００３３】

図３あるいは図４の表形式のデータ表示においては、表の個々のセルが一つのタンパク質と低分子化合物の相互作用に対応している。これをここでは「個々データ表示形式」と呼ぶ。しかし、個々データ表示形式においてはタンパク質の数や低分子化合物の数が増えるにしたがって、表のサイズが大きくなり、データ全体の把握が難しくなってくるという欠点がある。すなわち、データ数の増大に応じて表の個々のセルのサイズを変えなければ

、表全体が画面に入りなくなり、データ全体の様子を一望することができなくなる。逆に、表の個々のセルのサイズを小さくすることによって、表全体を画面内に収めるようにすると、セルに表示された相互作用データのパターンが細かくなり、その特徴の認識が困難になる。そこで、データ数が増大した場合も一望して表全体の相互作用パターンを認識可能にするために、図３あるいは図４における個々のクラスターを表上の一つのセルとして情報を表示することを可能にした。これをここでは「クラスター表示形式」と呼ぶ。

【００３４】

図５において、クラスター表示形式での情報表示例を示す。ラベル５０１にはクラスターの番号が入り、特徴量としてはクラスターに属する要素の数５０２と、クラスターに属する要素のリスト５０３が示されている。マトリクス部分５０４にはクラスターごとの測定データの平均値が色の濃さによって表示され、クラスターを構成する要素の数が数値によって示されている。個々データ表示形式による情報表示とクラスター表示形式による情報表示の切り替えが可能である。また、一つの表示形式における行や列の並べ替え、削除などの操作はもう一つの表示形式に反映される。クラスター表示形式においては、似たタンパク質同士、似た低分子化合物同士がクラスターを形成することから、代表的なデータを取りこぼすことなく可視化することができる。それと同時にクラスターの数を調節することによって、相互作用データの数が多いときも表示される表の行数、列数をコントロールできる。

【００３５】

個々データ表示形式とクラスター表示形式に相補的な情報表示形式として、「統計量表示形式」がある。これはデータの全部または一部に対して平均値、標準偏差などの統計計算を行い表示したり、異なるデータソースから抽出されたデータの件数を表示したりする形式である。統計量表示形式においては、相互作用データの数にかかわりなく、データの全体像を把握することができる。特に、データ数が増大した場合には、クラスター表示形式においても、一望して表全体の相互作用パターンを認識することが困難になってくる。このような場合に、統計量表示形式は、データの全体像を把握するという観点で非常に有効である。

【００３６】

本発明においては、表示形式を複数用意すると同時に、行列の各セル中に表示する情報として、要約の程度を変えたものを複数用意しておき、その中からデータ数に応じたものを選択して用いることを特徴としている。

【００３７】

タンパク質と低分子化合物の相互作用データの表示においては、４つの要約度（０－４）を用意する。要約度０では、データベースに格納されている情報や、そこから計算された統計量などをもれなく表示する。要約度１では、一つのセル当たり６４文字までの文字データ、記号、色彩を表示できる。データベース中のテキストフィールドで６４文字以下のものや、たとえ長いものであっても６４文字以下に情報を削減できるものであれば表示可能である。要約度２では、一つのセル当たり８文字までの文字データ、記号、色彩を表示できる。要約度３では、文字データは表示しない。全ての情報を色彩で表現する。

【００３８】

実装においては、要約度０における情報表示はフリーフォーマットとし、要約度１では一つのセルのサイズを縦６０ピクセル×横１２０ピクセルとして、その中に１６文字×４行分のテキストを表示する領域を確保する。要約度２では一つのセルのサイズを縦２０ピクセル×横６０ピクセルとして、その中に８文字×１行分のテキストを表示する領域を確保する。要約度３では一つのセルのサイズを縦５ピクセル×横５ピクセルとした。原理的には一つのセルのサイズを最低１ピクセル×１ピクセルにまで縮小することは可能であるが、マウスを使って個々のデータを操作可能なセルサイズを選択している。

【００３９】

これら４つの要約度における画面表示は、切り替え表示が可能である。図６に個々データ表示形式での４つの要約度別の情報の画面表示例を示す。

【００４０】
要約度０における画面表示６０１では、相互作用のデータ、低分子化合物のデータ、タンパク質のデータが詳細に表示されている。表示フォーマットは自由であり、タンパク質や低分子化合物の構造なども表示し操作することが可能である。
【００４１】
要約度１における画面表示６０２では、タンパク質関連の各種外部データベースへアクセスするためのキー、低分子化合物の名前や薬効、また相互作用の測定データの詳細な数値などを表示している。
【００４２】
要約度２における画面表示６０３では、表示される文字データは８文字までに限られるので、行や列を識別するためのラベルや、相互作用の測定データの主要な値などの限られた情報を表示している。
【００４３】
要約度３における画面表示６０４では、各セルがとる値を色彩情報に変換して表示している。これによって類似したデータを色彩のパターンから視認することができる。
【００４４】
選択されたデータ項目について、要約度によってどのように情報を要約するのかに関してルールを作る必要がある。基本的なルールは、要約度０においては、すべての情報の表示、要約度１と２においては文字の長さに応じた情報表示、要約度３においては色彩表示となっている。この基本的なルールにのっとり、詳細な要約のルールを、データベースに存在するそれぞれのデータ項目について定義する必要がある。
【００４５】
図７に、一例として、低分子化合物特徴テーブルについての要約ルール決定表を示す。要約度７０１に応じて、テーブル中のフィールドのうちどのデータ項目７０２を、どの場所７０３に、どのような要約ルール７０４で加工して画面表示をするかについての情報が与えられている。
【００４６】
フィールド名が要約ルール決定表に現れない場合は、そのフィールドは表示されないことを意味する。要約ルールが「そのまま」７０５の場合、データベースに格納されているデータをそのまま表示する。別の例として「色彩（２００，３００，４００，５００）」７０６の場合、値が２００未満、２００以上３００未満、３００以上４００未満、４００以上５００未満、５００以上の五つのケースについて色分け表示をする。このような要約ルール決定表をデータベース中のそれぞれのテーブルについて持つ必要がある。
【００４７】
以上、３つのデータの表示形式と、４つのデータの要約度を説明した。これらを組み合わせることによって多種多様な角度からデータを可視化することが可能である。本発明は、ユーザーが見たい情報を選択すると、そのデータ数に応じて最適なデータの表示形式とデータの要約度を自動的に決定する機能に特徴がある。
【００４８】
データの表示形式とデータの要約度を自動決定するための入力データとして、タンパク質と低分子化合物の相互作用データの可視化の例においては、タンパク質の数P、低分子化合物の数C、タンパク質クラスターの数Pc、低分子化合物クラスターの数Cc、及び、画面上における情報表示領域のパラメターｘ（高さ）、ｙ（幅）が必要である。クラスターの種類が複数ある場合は初期設定として登録されているクラスターの数を使用する。
【００４９】
図８にデータの表示形式とデータの要約度を決定するためのルールを表形式で示す。条件８０１を上から順番に見ていき、条件を満たしたところで、その行に記述されている表示形式８０２と、要約度８０３を採用する。条件を満たさない場合は、次の行の条件を見る。ここで、G、R、Gc、Rcは図８中で定義された数値である。以下この表を説明する。
【００５０】

P×C（表示画面内のセル数に該当）が一定値（この場合は３）より小さい場合、個々データ表示で要約度0を用いる。

【００５１】

P×C＞３で、かつG≦１１　＆　R≦１１の場合は、列方向特徴量表示数と行方向特徴量表示数がともに１である場合、タンパク質の数P、低分子化合物の数C共に2以上で、９以下となる。この場合は、要約度１を用いるので、一つのセルのサイズが縦６０ピクセル×横１２０ピクセルとなり、縦４５０ピクセル×横９００ピクセルの情報表示領域においては、全データの表示サイズは、縦２４０ピクセル×横４８０ピクセル～縦６６０ピクセル×横１３２０ピクセルとなる。これは、情報表示領域全体の１．５×１．５倍以内のサイズである。

【００５２】

タンパク質の数P、低分子化合物の数Cが増大するに従って、図８に従い順次、要約度を２、３と大きくしていく。さらにP，C数が増大した場合、クラスター表示に切り替え、タンパク質クラスターの数Pcと低分子化合物クラスターの数Ccが増大するに従って、要約度を１、２、３と増加させていく。

【００５３】

以上示した表示形式と要約度の切り替えを行うためのG、R、Gc、Rcに対する条件としては、全データの表示サイズが、情報表示領域全体の１．５×１．５倍以内のサイズになるような条件を設定している。データ表示領域のn×m倍以内に全データの情報を表示するという一般化された基準を満たすようにするには、

x×n≦P（又はPc）and y×m≦C（又はCc）

という一般化された条件を、データの表示形式と要約度の決定に用いればよい。

【００５４】

このようにすることによって、データの全体、あるいはその一定の倍数のデータ量を、情報表示領域内で表示することが可能になり、かつ、データ数の増減に応じて要約度を上下させることによって、セル内に、一望して認識可能でかつ最大限の情報量を表示可能になる。これにより、表示すべきデータ数にかかわらず、個別セル内から得られる情報量を最大に保ちつつ、データの全体像の観察が可能になる。

【００５５】

新規創薬ターゲットの発見のプロセスにおいては、タンパク質と低分子化合物の相互作用を可視化すると同時に、他の関連する生体関連の相互作用についても同時に情報を得て、包括的に情報を整理し、理解することが極めて重要である。関連する生体関連の相互作用の例として、低分子化合物同士の薬効や毒性に関する相互作用、タンパク質同士の相互作用、タンパク質と発現に関する情報などが挙げられる。本発明においては、これら関連情報を取得し、取得したデータ数に応じて、上述した表示形式と要約度の決定ルールに従い、表示することが可能である。

【００５６】

関連情報の取得は、以下のように行う。表示されているデータテーブル内の着目するセル領域を選択し、このセル領域に属する低分子化合物ＩＤとタンパク質ＩＤを抽出する。これらのＩＤを、関連データテーブル中で検索し、検索されたＩＤに付随する情報を関連データテーブルから抽出する。

【００５７】

図９に、関連情報抽出の具体的な方法を示す。タンパク質－低分子化合物相互作用テーブル９０１のうち（C5,P12）と（C9,P12）の二つに着目しているとき、タンパク質間の結合強度を１００を最大値として規格化したタンパク質－タンパク質相互作用テーブル９０２と、発現ライブラリーにおける定性的なタンパク質の発現量を示すタンパク質－発現テーブル９０３からはタンパク質のIDがP12であるもののうち、データが存在するものを抽出する。同様に低分子化合物間の多剤併用による効果のある・なしのデータを格納した低分子化合物－低分子化合物相互作用テーブル９０４からはIDとしてC5,C9を持つもののうち、データが存在するものを抽出する。

【００５８】
関連情報の抽出結果は図１０のように、抽出元の表ごとに整理されて表示される。ユーザーが見たい表を選択すると、そのヒット件数に応じて自動的に情報の表示形式と要約度が設定され、設定された表示形式と要約度で情報が画面表示される。そのようにして表示された情報の一部から、また関連情報を取得することができる。したがって、本発明によって多次元の相互作用データを1対1相互作用データ間のリンクを効率的にたどることで可視化することができる。

【００５９】
本発明の可視化方法を実装したインターフェースにおいては、画面表示された情報のうち一部を選択し、選択されたデータに対して、複数のアクションから選択したアクションを実施し、アクションの結果得られた情報が画面表示される。図１１にユーザーインターフェースの例を示す。表示モードの変更ボタン１１０１、要約度の変更ボタン１１０２、関連情報取得ボタン１１０３に加え、行や列の入れ替え、並べ替え、クラスタリング、削除などのアクションに関連する機能群１１０４と、特徴的な行や列、代表的なサブセットとしての行や列などの選択に関連する機能群１１０５を備える。また、画面上に表形式で表されているセルの一つ一つに対してマウス操作によるアクションが割り当てられていて、それによって、行や列を選択したり、関連情報表示画面１１０６にセルの中には表示できない長い文字列データなども表示したりできる。

【図面の簡単な説明】

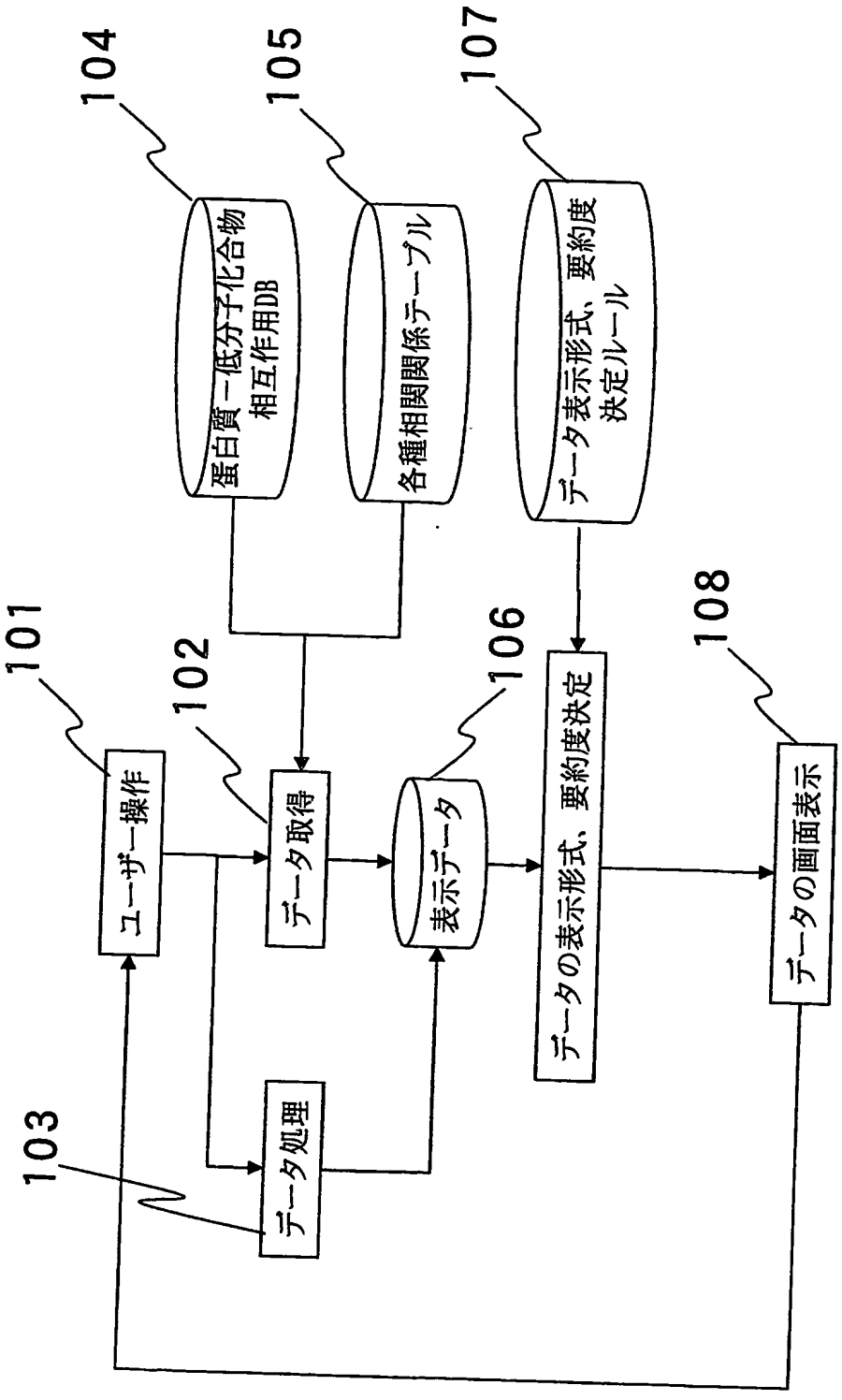
【００６０】
【図１】データ可視化のフローチャート。
【図２】低分子化合物とタンパク質の相互作用データの画面表示例。
【図３】相互作用データプロファイルを用いたクラスタリング結果に基づいてソートされたデータの画面表示例。
【図４】行および列の特徴量を用いたクラスタリング結果に基づいてソートされたデータの画面表示例。
【図５】クラスター表示形式での情報表示例。
【図６】個々データ表示形式での４つの要約度別の情報の画面表示例。
【図７】データの表示形式とデータの要約度を決定するためのルール。
【図８】低分子化合物物性テーブルについての要約ルール決定表。
【図９】関連情報抽出方法の概要。
【図１０】関連情報の抽出結果。
【図１１】本発明を実装したユーザーインターフェースの画面例。

【符号の説明】

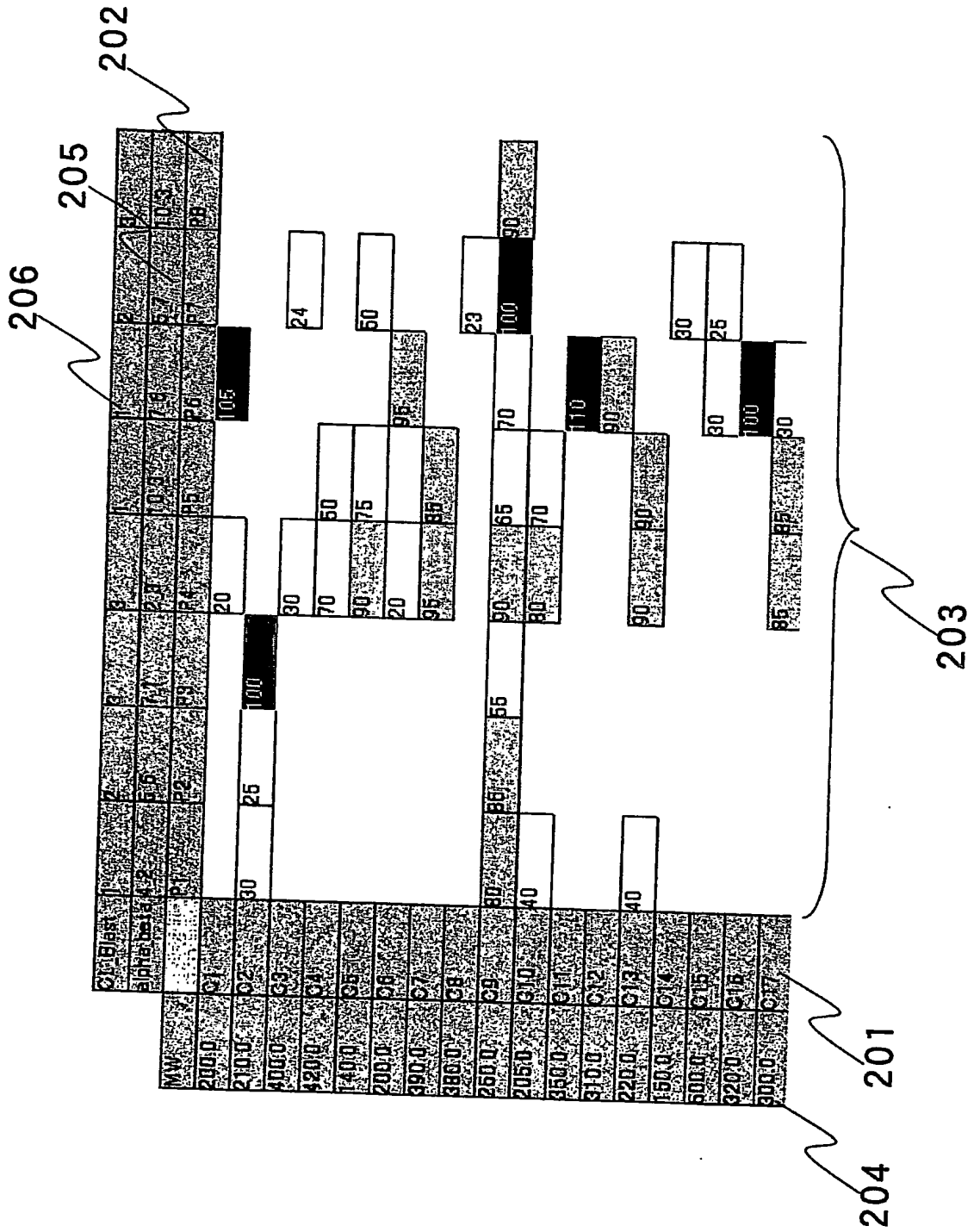
【００６１】
１０１：ユーザー操作、１０２：データ取得、１０３：データ処理、１０４：蛋白質－低分子化合物相互作用データベース、１０５：各種相関関係テーブル、１０６：表示データ、１０７：データ表示形式と要約度決定ルール。
２０１：低分子化合物のラベル、２０２：タンパク質のラベル、２０３：マトリクス部分、２０４：分子量、２０５：アルファヘリックスとベータストランドの数、２０６：相同性に基づくクラスタリング情報。
３０１：低分子化合物クラスターA、３０２：低分子化合物クラスターB、３０３：低分子化合物クラスターC、３０４：タンパク質クラスターA、３０５：タンパク質クラスターB、３０６：タンパク質クラスターC、３０７：特定の低分子化合物とタンパク質の組からなるクラスター、３０８：一つのタンパク質について特異的に相互作用をもつ化合物の組からなるクラスター。
４０１：分子量の比較的大きなクラスターA、４０２：中程度の分子量を持つクラスターB、４０３：分子量の比較的小さなクラスターC、４０４：アミノ酸配列の相同性に基づいてクラスター１、４０５：アミノ酸配列の相同性に基づいてクラスター２、４０６：比較的相互作用が高い領域。

５０１：ラベル、５０２：クラスターに属する要素の数、５０３：クラスターに属する要素のリスト、５０４：マトリクス部分。
６０１：要約度０における画面表示、６０２：要約度１における画面表示、６０３：要約度２における画面表示、６０４：要約度３における画面表示。
７０１：要約度、７０２：データ項目、７０３：場所、７０４：要約ルール、７０５：ルール「そのまま」、７０６：ルール「色彩（２００，３００，４００，５００）」。
８０１：条件、８０２：表示形式、８０３：要約度。
９０１：タンパク質－低分子化合物相互作用テーブル、９０２：タンパク質－タンパク質相互作用テーブル、９０３：タンパク質－発現テーブル、９０４：低分子化合物－低分子化合物相互作用テーブル。
１１０１：表示モードの変更ボタン、１１０２：要約度の変更ボタン、１１０３：関連情報取得ボタン、１１０４：アクションに関連する機能群、１１０５：選択に関連する機能群、１１０６：関連情報表示画面。

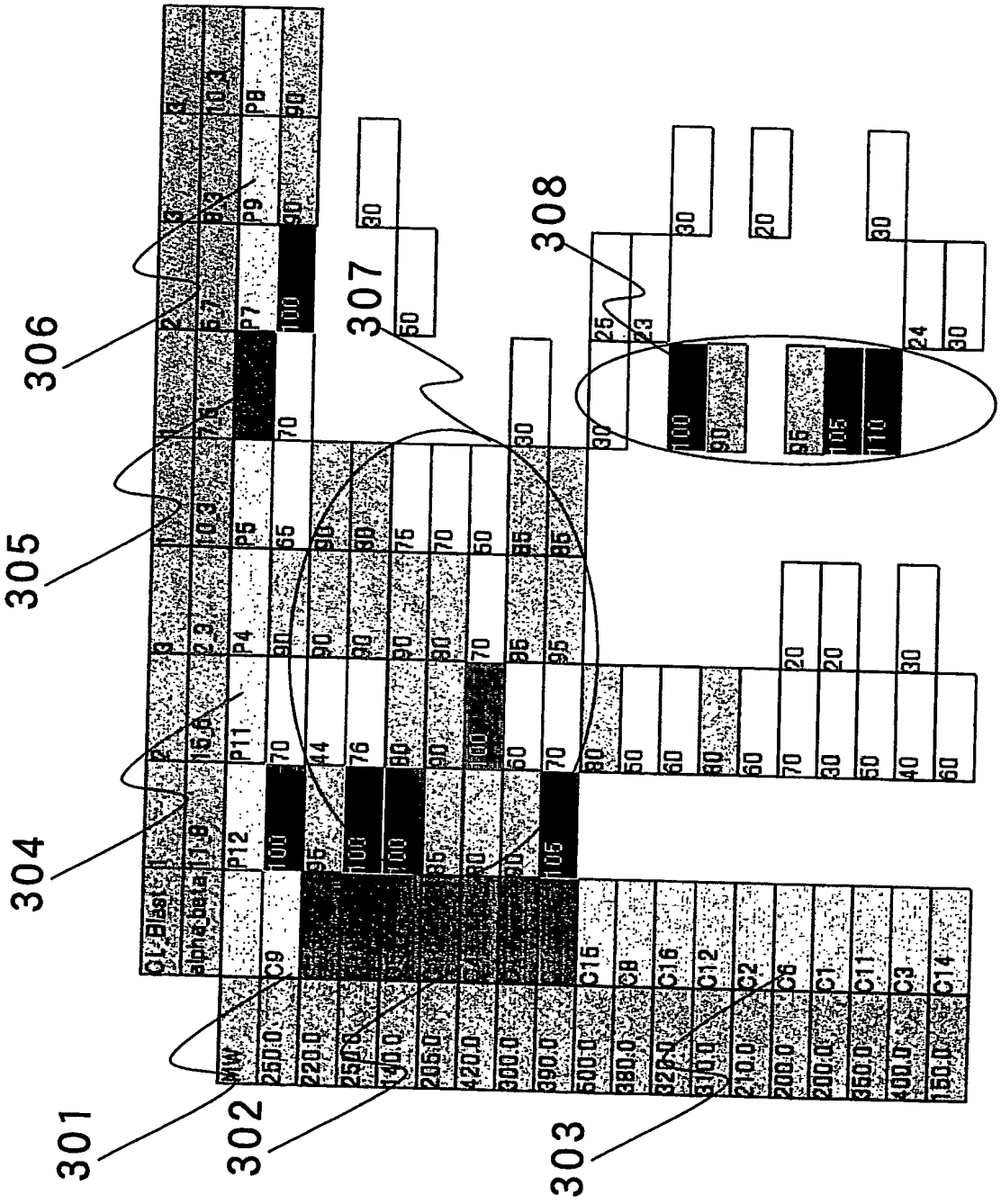
【書類名】図面
【図１】


104
105
107
蛋白質－低分子化合物
相互作用DB
各種相関関係テーブル
データ表示形式、要約度
決定ルール
101
102
106
108
ユーザー操作
データ取得
表示データ
データの表示形式、要約度決定
データの画面表示
103
データ処理

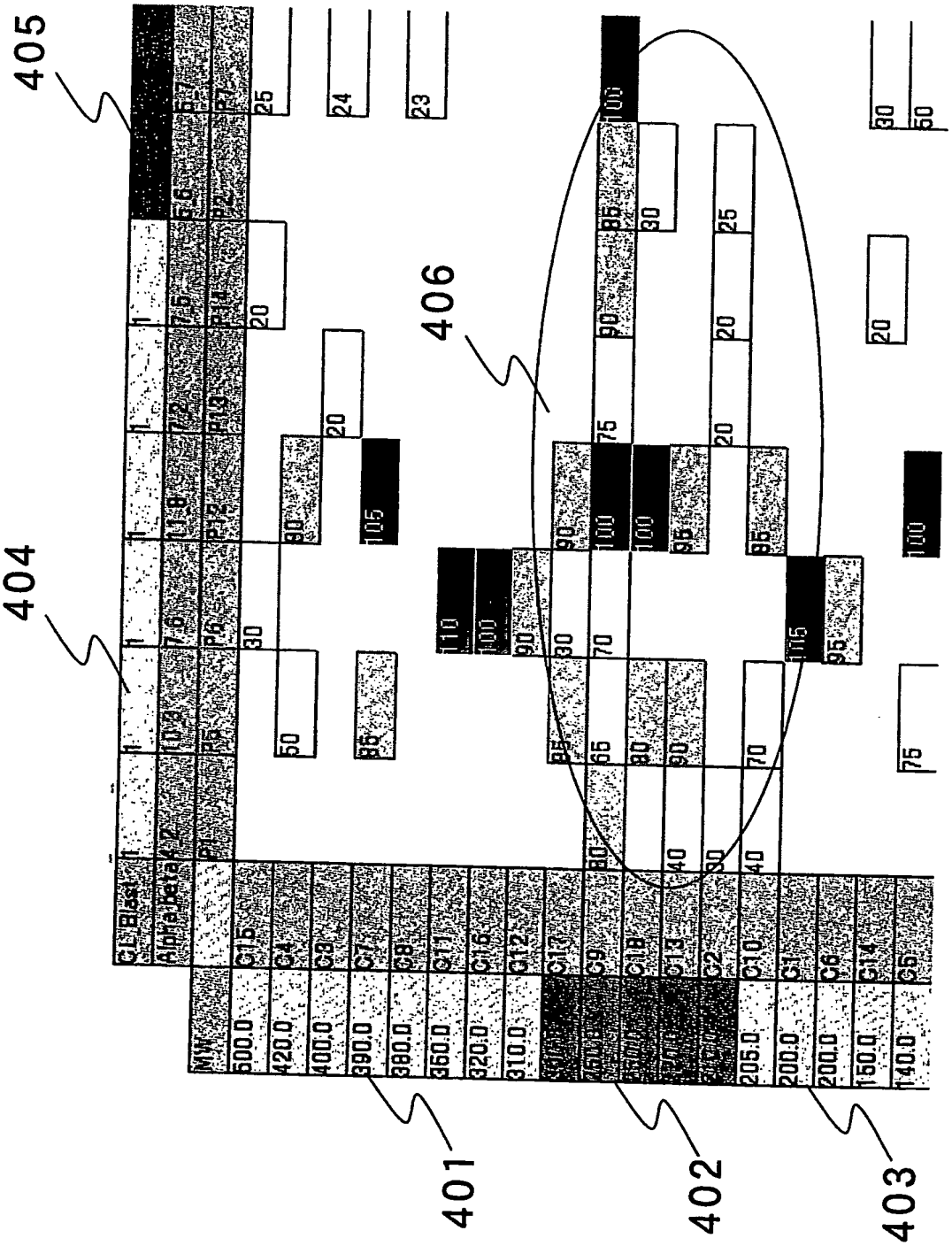
【図２】



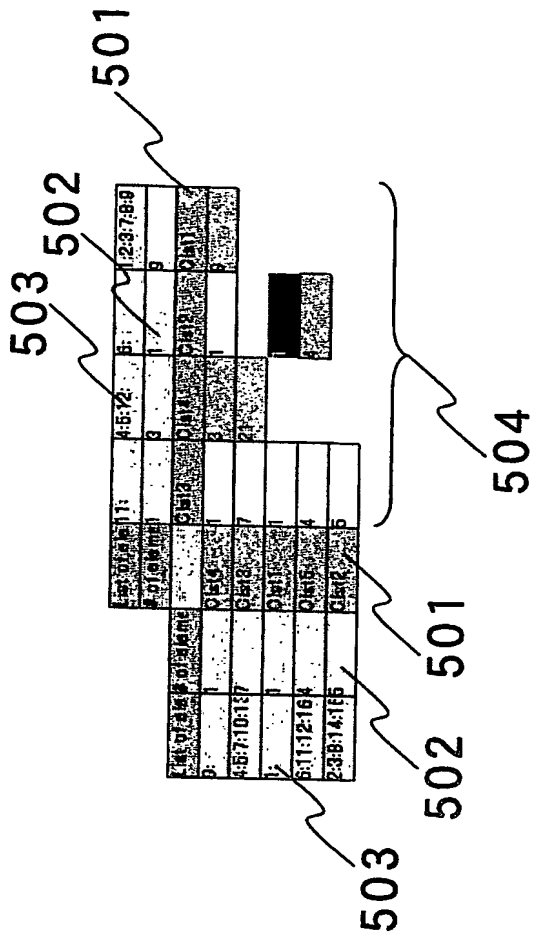
【図３】



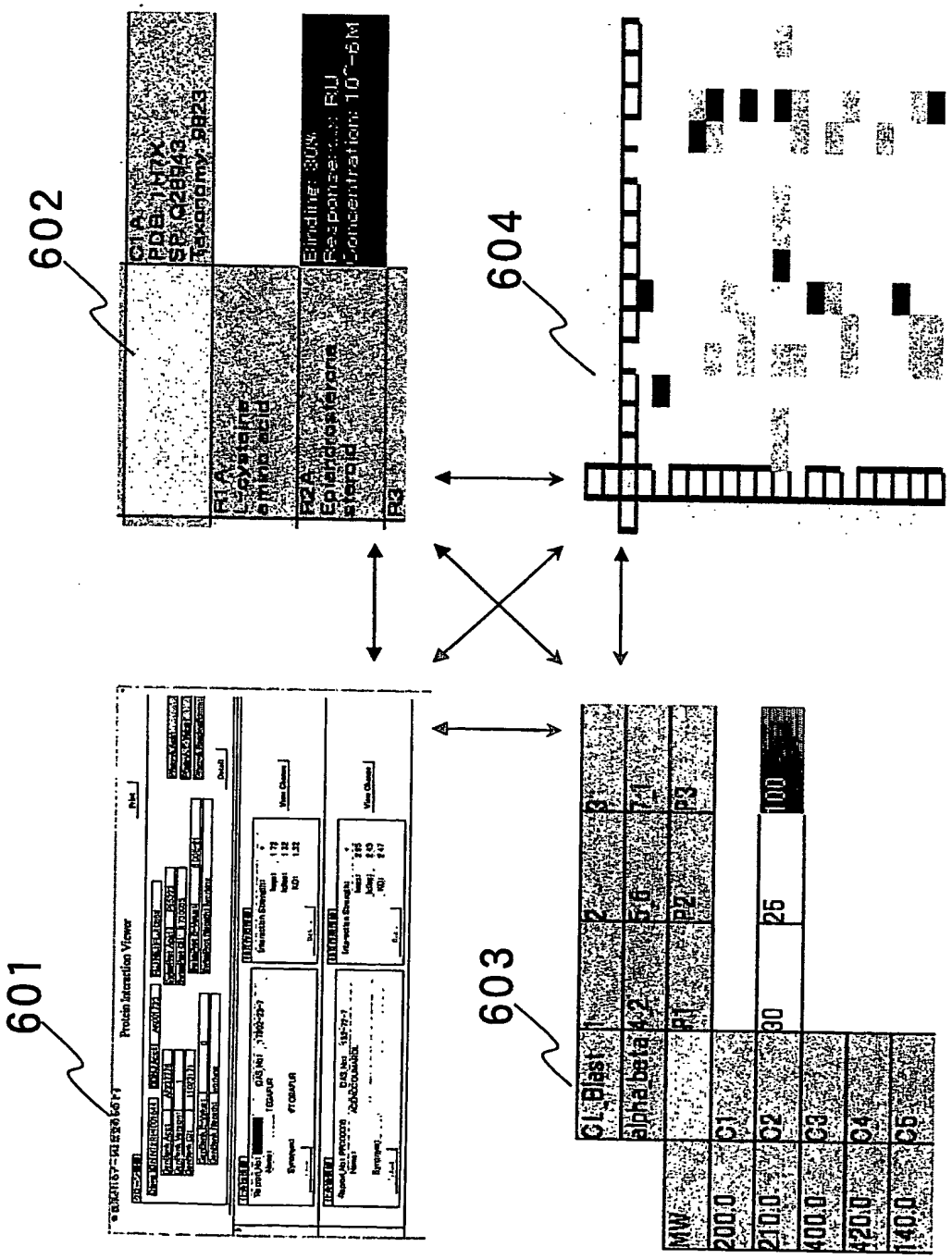
【図４】


405
406
404
401
402
403
Cl Blast
Alpha beta
MW
600.0
420.0
400.0
390.0
380.0
360.0
320.0
310.0
205.0
200.0
200.0
150.0
140.0

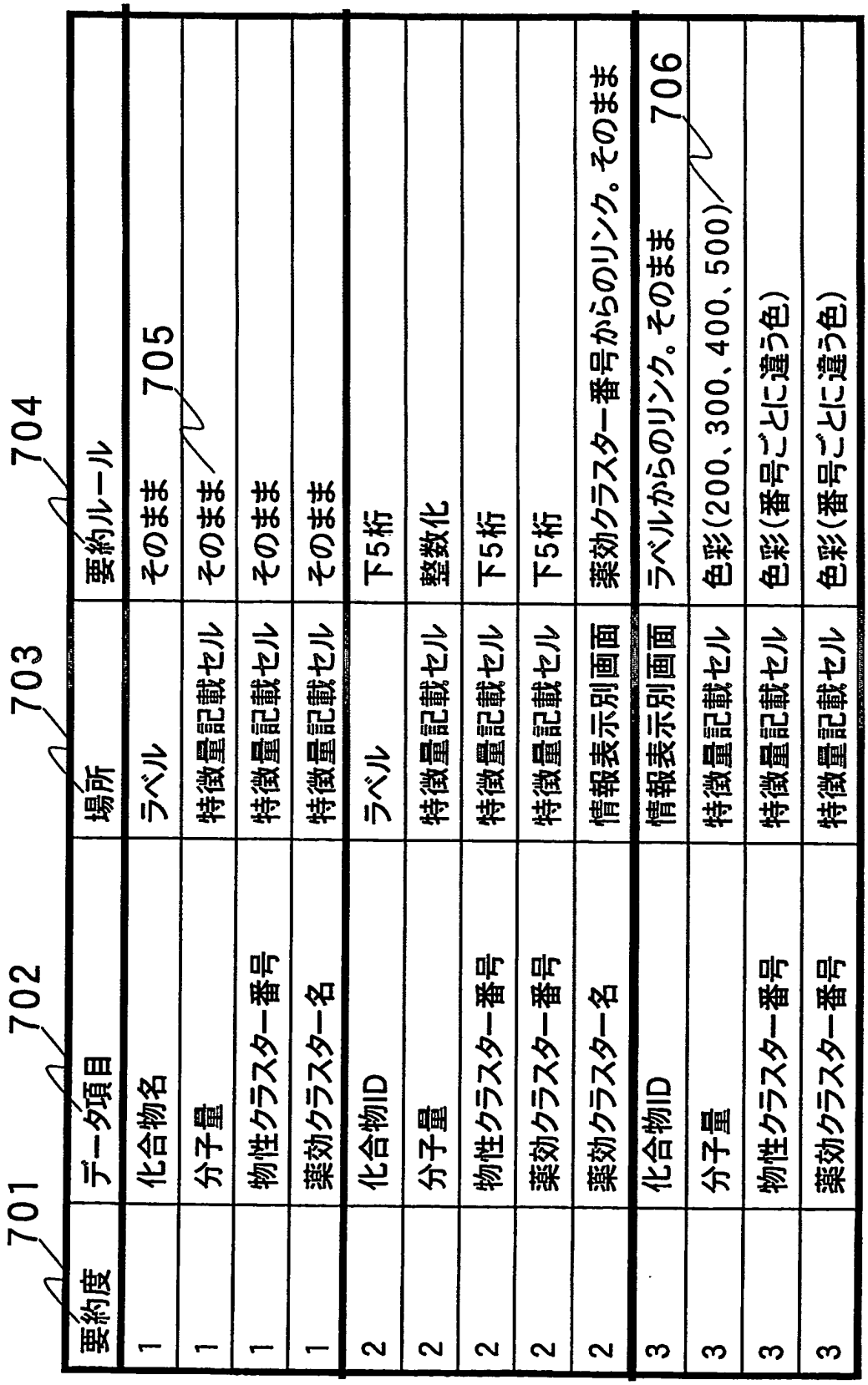
【図５】



【図６】


602
604
601
603
Binding: 30%
Response: RU
Concentration: 10^-6M
R1 Alanine amino acid
R2 Epiandrosterone steroid
R3
Protein Interaction Viewer
GL Blast
alpha beta
MW
200.0
210.0
400.0
420.0
140.0
C1
C2
C3
C4
C5
30
25
100

【図７】



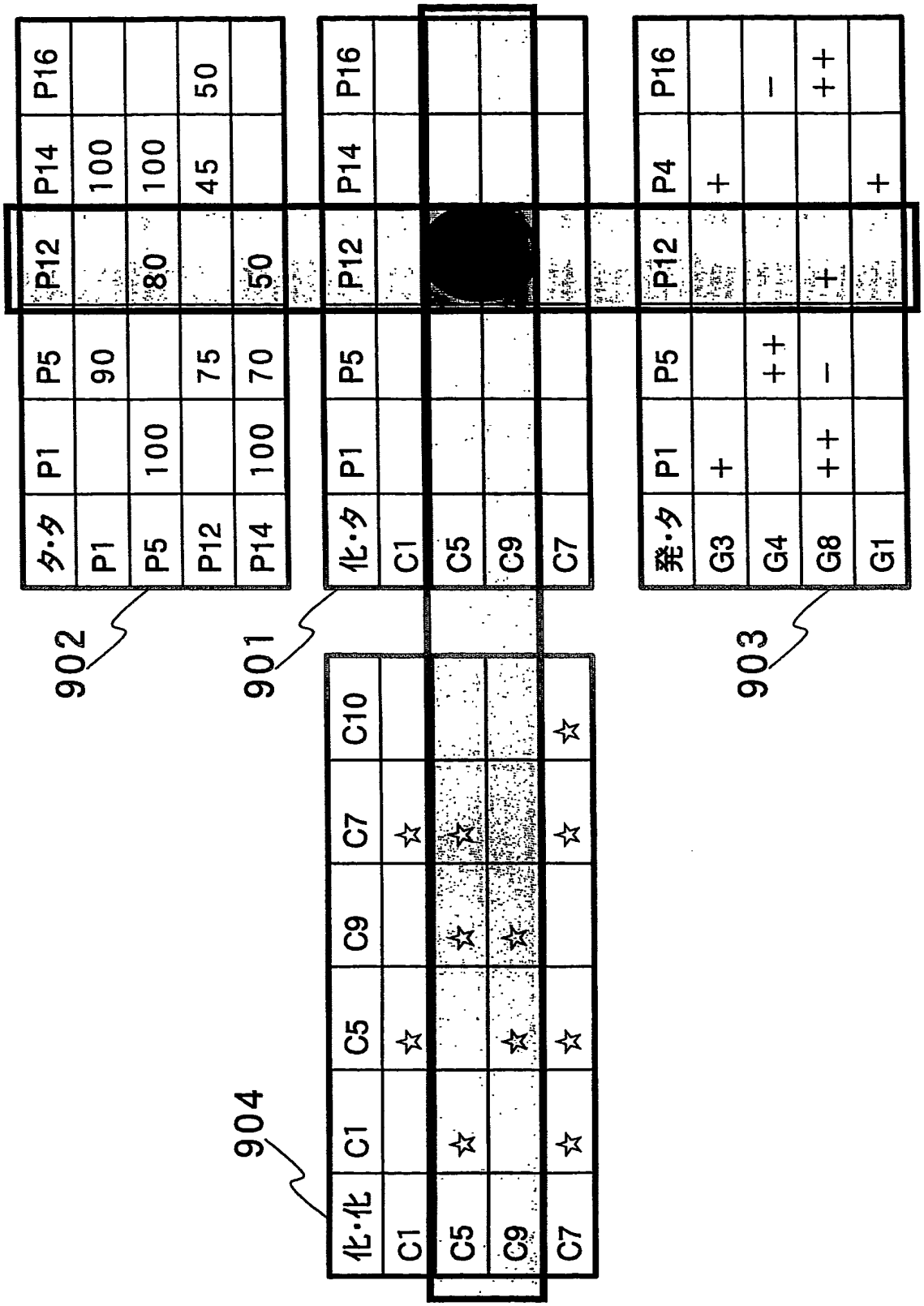
要約度	データ項目	場所	要約ルール
1	化合物名	ラベル	そのまま
1	分子量	特徴量記載セル	そのまま
1	物性クラスター番号	特徴量記載セル	そのまま
1	薬効クラスター名	特徴量記載セル	そのまま
2	化合物ID	ラベル	下5桁
2	分子量	特徴量記載セル	整数化
2	物性クラスター番号	特徴量記載セル	下5桁
2	薬効クラスター番号	特徴量記載セル	下5桁
2	薬効クラスター名	情報表示別画面	薬効クラスター番号からのリンク。そのまま
3	化合物ID	情報表示別画面	ラベルからのリンク。そのまま
3	分子量	特徴量記載セル	色彩(200、300、400、500)
3	物性クラスター番号	特徴量記載セル	色彩(番号ごとに違う色)
3	薬効クラスター番号	特徴量記載セル	色彩(番号ごとに違う色)

【図８】

801 条件	802 表示形式	803 要約度
P×C≦3	個々データ表示	0
G≦11 & R≦11	個々データ表示	1
G≦34 & R≦22	個々データ表示	2
G≦135 & R≦270	個々データ表示	3
Gc≦11 & Rc≦11	クラスター表示	1
Gc≦34 & Rc≦22	クラスター表示	2
Gc≦135 & Rc≦270	クラスター表示	3
上記以外	統計量表示	0

G: P+列方向物性表示数+1
R: C+行方向物性表示数+1
Gc: Pc+列方向物性表示数+1
Rc: Cc+行方向物性表示数+1

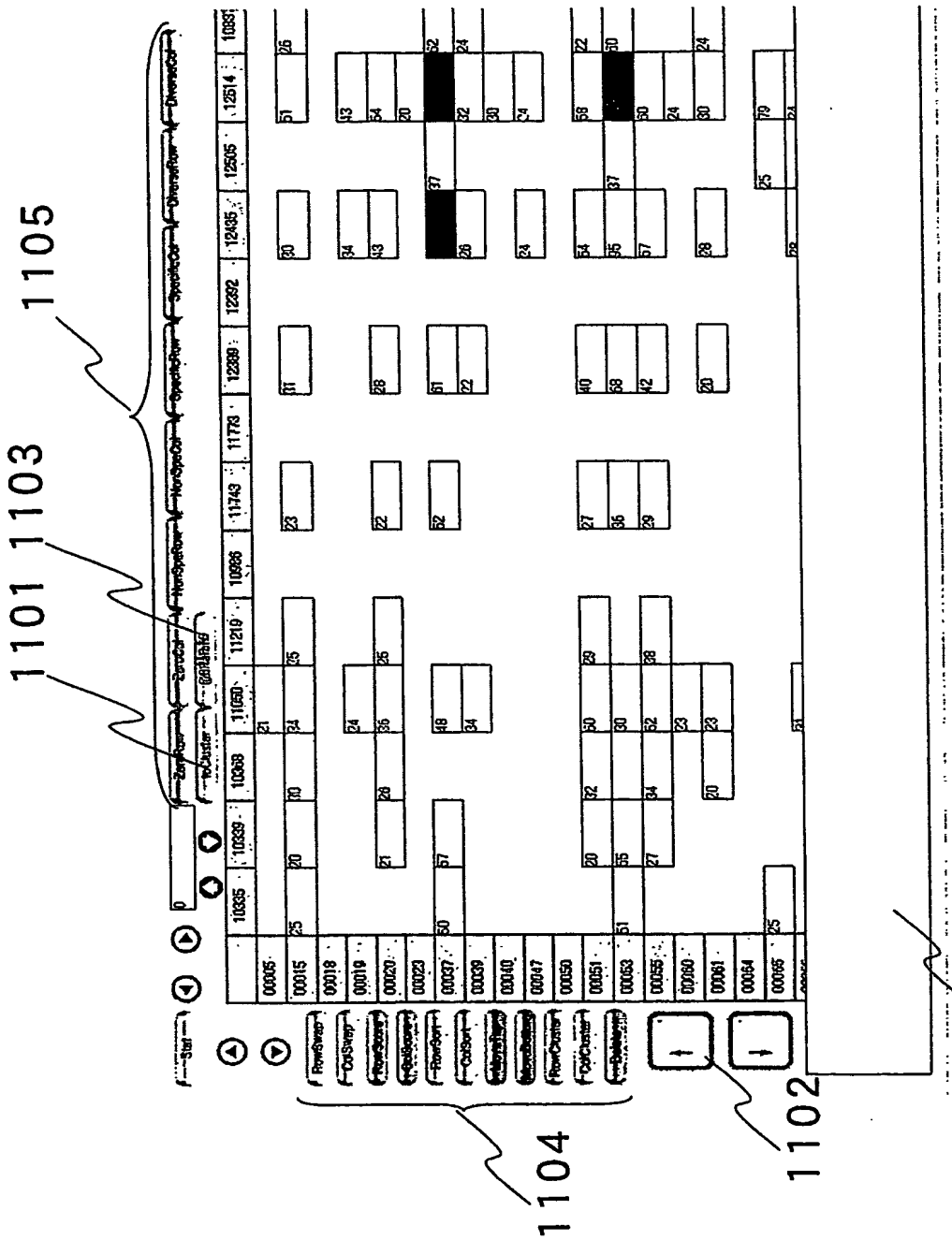
【図９】


902
901
903
904
タ・タ
化・タ
発・タ
化・化
P1
P5
P12
P14
P16
C1
C5
C9
C7
C10
G3
G4
G8
G1
P4
90
100
80
75
50
70
45
＋
－
＋＋
☆

【図１０】

(R5,C12), (R9, C12)の関連情報
タンパク質―発現テーブルより：　xx件
タンパク質―タンパク質相互作用テーブルより：　yy件
低分子化合物―低分子化合物相互作用テーブルより：　zz件

【図１１】



【書類名】要約書
【要約】
【課題】　二つの事象間の相関データを行列形式で表示する可視化方法において、相関データの全体としての観察と少数データの詳細な観察を交互に繰り返す作業の手間を軽減する。
【解決手段】　データ数の規模の変動に応じて、相関データの単位が異なる予め用意された複数のデータ表示形式の中から一つを自動的に選択し、また、個々のセルに関する情報（相関データや各事象に関する情報）について要約度の異なる予め用意された複数の表示方法の中から一つを自動的に選択して、相関データと個々のセルに関する情報を表示する。
【選択図】　図１

特願２００３－３４８４３８

出　願　人　履　歴　情　報

識別番号　　[５０１２６００８２]

１．変更年月日　　２００２年１２月１３日
[変更理由]　　住所変更
住　所　　千葉県木更津市かずさ鎌足２丁目６番地７
氏　名　　株式会社リバース・プロテオミクス研究所